Web usage mining at an academic health sciences library: an exploratory study

By Paul J. Bracke, MS
paul@ahsl.arizona.edu
Head of Systems and Networking

Arizona Health Sciences Library
1501 North Campbell Avenue
P.O. Box 245079
Tucson, Arizona 85724-5079

Objectives: This paper explores the potential of multinomial logistic regression analysis to perform Web usage mining for an academic health sciences library Website.

Methods: Usage of database-driven resource gateway pages was logged for a six-month period, including information about users' network addresses, referring uniform resource locators (URLs), and types of resource accessed.

Results: It was found that referring URL did vary significantly by two factors: whether a user was on-campus and what type of resource was accessed.

Conclusions: Although the data available for analysis are limited by the nature of the Web and concerns for privacy, this method demonstrates the potential for gaining insight into Web usage that supplements Web log analysis. It can be used to improve the design of static and dynamic Websites today and could be used in the design of more advanced Web systems in the future.

INTRODUCTION

The Web has had a significant impact on libraries, changing the formats and methods in which they present their resources and services to users. Most libraries use a Website as a tool for organizing and providing access to resources in print and electronic formats. The impact of the Web for libraries, however, has been more profound than providing a new access point for its users. The Web has also changed the information-seeking behavior of and information use by library users and perhaps expanded the definition of who libraries' users are.

In the not-so-distant past, searchable bibliographic databases were limited in their availability, accessible only to information professionals or from workstations located in the library. The Web has changed this, allowing information discovery tools such as Alta Vista, Yahoo, and Google to become a part of everyday life for many. The increased availability of search tools seems to have changed the behavior of users, many of whom no longer need to come to the library as frequently as before. This change in the use of physical libraries is further demonstrated by a decline in foot traffic and the number of reference questions fielded by libraries in the past five years [1].

Despite the declining in-person usage of libraries, many users still depend on libraries to aggregate and provide access to a range of electronic resources. In the past, gate counts and reference and reshelving statistics were used to measure the use of physical resources. Assessment continues to be important in an electronic environment to determine the effectiveness of a library in shepherding its users to appropriate resources. In addition to the need for understanding how a library's own users access its resources, assessment also creates possibilities for identifying new markets for resources and services. The Web makes libraries virtually available to users who may not be physically able to access them. For some libraries, the Web also provides an opportunity to identify new services to better serve existing users, to possibly redefine who those users are, and to reach out to those who have not been served before.

This paper will explore the potential of a Web usage mining technique, regression analysis, to analyze navigational routes used to access the gateway pages of the Arizona Health Sciences Library (AHSL) Website over a six-month period.

LITERATURE REVIEW

A number of articles have discussed Web server log analysis for libraries, since libraries began to develop

Web presences [2–9]. These articles describe summary level metrics of Website usage, such as the total number of user sessions, broken down by variables such as date, time, or host domain of the requestor. As noted by many of these authors and by Goldberg [10], these studies have been constrained by two main factors. One, data provided by the hypertext transfer protocol (HTTP) that governs user transactions on the Web are very limited. Second, usage logs are designed for use by system administrators, not for tracking users. While the information available in logs is limited, some user and resource usage data can be gleaned from them. For example, the Internet protocol (IP) or network address of users can provide some insight into who is using a site, and the requested file can be used to make some conclusions about what content is being used.

Statistical analysis of Web data is well developed in data mining and its subspecialty, Web usage mining. The literature on data mining contains descriptions of sophisticated statistical analyses of Website usage, with applications including personalization and system improvement [11]. These analyses apply a variety of techniques including ordinary least squares (OLS) and logistic regression, cluster analysis, decision trees, and neural networks. Web usage mining often analyzes sequences of page accesses to provide personalization and targeted marketing [12, 13]. Feng and Murtagh provide an example of using Web usage mining techniques to develop a personalization system [14]. Davis analyzes the information-seeking behavior of chemists based on Web log analysis [15].

The data mining literature tends to focus on business applications in which the purpose of a Website is to maximize revenue and provide opportunities for sales with advertising as focused as possible. Libraries, on the other hand, support a range of information needs, so the purpose of usage analysis is to gauge how well user needs are being met. Analysis is complicated by the impossibility of identifying user intent or need through Web server logs and the lack of a one-to-one relationship between need and resource. A single information resource may support many needs, yet all of these needs would be represented as identical uses in log files. Libraries can uncover basic patterns in the use of information sources and correlate it with informal knowledge about information needs in a library or with qualitative data gathered through surveys or focus groups. A recent issue of *Information Technology and Libraries* describes a number of efforts to engage in data mining to improve information access [16–19].

DATA

AHSL uses a series of gateway pages as intermediate screens between an electronic resource and other access points, such as records in ALOE, the library catalog, or an alphabetical list of electronic journals. Generally, each resource has a single gateway page, which is dynamically generated from a database of electronic resources. The database and scheme of gateway pages simplify Website management and make integrating proxy links for remote access easier. Gateway pages are created for freely available journals and databases, such as PubMed, as well as resources licensed by AHSL with access restrictions. Licensed resources are typically open only to users affiliated with the institutions served by AHSL: The University of Arizona (UA), University Medical Center (UMC), and University Physicians, Inc. (UPI). Users who access the gateway page from an IP address not associated with one of these institutions will be unable to access the resource, unless they can be authenticated as valid members of the AHSL user base. Because they provide access to online library collections, AHSL considers these gateway pages the core of the Website.

The gateway pages can also be used to capture usage data. These pages are available to any user on the Web, although the resources to which they lead may not be open to all users. When a gateway page is requested, information about the request may be logged to a database, independent of the normal Web server logging process. This information includes the HTTP variables normally available to a Web server. Additionally, information about the requested resource (e.g., format) may also be pulled from the database and logged as the gateway page is generated. Unfortunately, information about user affiliation is unavailable for logging and analysis, in part due to the technical details of UA's authentication system and the heterogeneous nature of AHSL's user base. Valid AHSL users include non-UA affiliates, such as hospital affiliates.

The resulting data set has a limited number of usable variables, but they are the most detailed that can be gathered given the technical limitations of the Web and the UA computing environment. The following variables were collected from September 2002 through March 2003.

1. User Type (UAIP): User type is a dichotomous variable coded from the user's IP address. Users are coded as affiliated (noted by a 1) if they have a UA, UMC, or UPI IP address and are considered unaffiliated (noted by a 0) otherwise. If directory-based authentication were available and required to access a gateway page, it would also be possible to capture information about departmental affiliations and status in an institution (e.g., graduate student, faculty, or staff). This measure is limited, because not all users logged as unaffiliated in this study are actually unaffiliated. Remote access is provided to affiliated, off-campus users who are able to be successfully authenticated and use the AHSL proxy server.
2. Resource Type (TYPEID): Resource type captures the format of the resource described by the requested gateway page. This variable has four categories (book, journal, database, and Internet resource). Observations with Internet resource as resource type are dropped because of errors in data collection.
3. Restriction Type (ACCID): Restriction type is a categorical variable with three categories (open, UA only, and AHSL only). It indicates the type of restriction (or lack thereof) dictated by licensing agreements. Open

access materials are open to any user, affiliated or not. It is recoded into a dummy variable, UAONLY, with a value of 1 if ACCID equaled UA only or AHSL only.

4. Referrer Type (REFID): Referrer type is a categorical variable with five categories (AHSL Website, UA Main Library Website, other UA Website, Internet search engine, and other Website). One piece of data often transmitted through HTTP is referrer, the page from which a user is linked. This information is useful, because it indicates how users arrived at a particular page. It is also problematic, because it is not always available. Referrers may be unavailable for a variety of reasons, including the use of a bookmark or favorites list or the presence of some security devices or software. Of the 354,886 observations in the data set, 225,512 (63.5%) have a valid referrer.

5. AHSL Website Usage Mode (AHSLMODE): AHSL Website usage mode consists of four categories (catalog search, Website browse, Website search, and AZHIN). This variable describes the navigation method used to arrive at the gateway page in the AHSL Website.

Even with the expanded logging used to generate this data set, limitations to the data make characterizing the usage of gateway pages difficult. To understand the usage of these pages, it is necessary to understand how users arrive at them. In particular, it is important to determine whether navigational route varies by information need. While the data set contains no information about user intent or about the questions that users seek to answer when they visit the site, it does contain information about the resources that the users examine in the process of satisfying their needs.

The TYPEID variable in the raw dataset records information about the type of resource (i.e., format) associated with each gateway page view. While information format is certainly not synonymous with information need, it is assumed in this analysis that different formats may be used to meet different needs, even if those needs are unknowable to the researcher. So, format will be used as a proxy measure for information need. The dummy variables JOURNAL and BOOK represent two of the three resource types in TYPEID, with database as the base category.

Some of these page views might have been part of an iterative process of resource selection, in which the user viewed the gateway pages and decided not to connect further. Unfortunately, it was not possible to log whether the user left the site, so it was assumed that the use of gateway pages led to the use of the resources to which they link. Additionally, as mentioned earlier, referring uniform resource locator (URL) was a limited variable for tracking navigational path.

METHODOLOGY

Because the variables in the data set are all unordered, categorical variables, multinomial logistic regression (MLR) analysis is used instead of OLS regression. MLR is an extension of logistic regression, which estimates the log odds of a dichotomous dependent variable having one of two values. The coefficients estimated for the independent variables in the model can then be used to transform the log odds into the probability of the dependent variable having one of two values. MLR extends this technique to estimate the expected probabilities when a dependent variable is a categorical variable, with no order in the categories. One category of the dependent variable is omitted from model estimation in MLR, with coefficients for each independent variable then estimated for the remaining categories of the dependent variable compared to the base category. These coefficients describe the effect of an independent variable on the probability that an observation will fall in the estimated category instead of the base category [20].

EFFECT OF RESOURCE TYPE ON NAVIGATIONAL ROUTE

Models

To answer the question of whether navigational route varies by type of sought resource, a series of models were estimated to determine whether the format of resource had an effect on which route to the gateway page was used. It could be expected that users looking for a known item, such as a journal or a textbook, would want to access that item as quickly as possible and thus be more likely to be referred to the site via a search engine than those seeking a tool for information discovery (i.e., a database). Additionally, because the AHSL Website was designed as the primary access point for these gateway pages, it was expected that it would be the top referrer, regardless of resource type. The results of these models are in Table 1. In each model, the AHSL Website was used as the base category of the referrer variable, the dependent variable.

Model 1 is a simple bivariate model in which referring URL (REFID) is regressed on whether the user was an affiliate of the University of Arizona (UAIP). This model verifies that type of user has a significant effect on navigational route for every category of referring URL. To test hypotheses about the effect of resource type, dummy variables for books (BOOK) and journals (JOURNAL) were constructed for model 2, and a variable representing the type of access restriction (UAONLY) was added. The added terms were significant in each category, with two exceptions. While parameter estimates for BOOK on UA Main Library and other Website were made, the standard error, z scores, and *P* values were not calculated, because each of these referrers had no cases of referring users to gateway pages for a book. These terms are significant in other categories of REFID, however, and a log-likelihood test indicates that model 2 fits the data better than model 1.

Models 3 and 4 test possible interactions between user status (UAIP) and both whether access is restricted (UAONLY) and whether a resources is a journal (JOURNAL). Some pages that link to the AHSL gate-

Table 1
Multinomial logistic regression (MLR) models of gateway page usage

Referrer comparison	Model 1	Model 2	Model 3	Model 4
UA Main Library/AHSL Website				
Constant	−3.771381*	−2.112803*	−2.61161*	−2.61463*
UAIP	0.9049907*	1.061856*	1.858378*	1.864771*
UAONLY		0.4526433*	1.336038*	1.309219*
BOOK		−50.97238**	−22.54961§	−51.27881**
JOURNAL		−7.249913*	−7.249781*	−6.31391*
UAIPXUAONLY			−1.339499*	−1.298764*
UAIPXJOURNAL				−2.240274*
Other UA Website/AHSL Website				
Constant	−4.118606*	−3.555309*	−3.77955*	−3.838584*
UAIP	−0.4297536*	−0.4068737*	0.2365569‡	0.3775304*
UAONLY		0.2606536*	0.5636248*	0.2363121†
BOOK		−1.283475*	−1.331699*	−1.143772*
JOURNAL		−0.9991079*	−1.030711*	−0.5793111*
UAIPXUAONLY			−0.7623108*	−0.2019624§
UAIPXJOURNAL				−0.8676843*
Search engine/AHSL Website				
Constant	0.5505684*	−0.0380126†	−0.0626743*	−0.0668061*
UAIP	−6.401237*	−6.416076*	−6.318304*	−6.129807*
UAONLY		−0.3880272*	−0.3381839*	−0.3464635*
BOOK		1.196881*	1.179413*	1.1884*
JOURNAL		1.04087*	1.018436*	1.031968*
UAIPXUAONLY			−0.1064646§	−0.0062529§
UAIPXJOURNAL				−0.3050727§
Other Website/AHSL Website				
Constant	−7.05658*	−6.885259*	−6.902591*	−6.942889*
UAIP	−3.987046*	−4.0013*	−23.14476*	−22.16871*
UAONLY		0.2027124*	0.2584144§	0.1063203§
BOOK		−44.91606**	−23.02093§	−44.75147**
JOURNAL		−0.3843273*	−0.4210721§	−0.2101326§
UAIPXUAONLY			19.22398**	20.55961**
UAIPXJOURNAL				−46.1379**
Pseudo R-Square	0.2669	0.3280	0.3290	0.3292
Log Likelihood	−134611.58	−123382.04	−123205.21	−123167.57

* $P < 0.01$.
† $P < 0.05$.
‡ $P < 0.10$.
§ $P > 0.10$.
** P not estimated.

way pages contain a note about the type of access restriction, so unaffiliated users would be less likely to click on a link labeled ''UA Only.'' It is thus plausible that the type of access restriction would have an effect on the routes users take to a gateway page. It also seems that users seeking particular types of resources might take a different route to a gateway page if they are affiliated with AHSL. For example, locating electronic journals is an area of emphasis for AHSL's educational program, so exposure to such training might make it more likely that seeking a journal affects navigational route for affiliated users differently than for unaffiliated users.

Each of these interaction terms was significant in at least one category, although their inclusion complicates interpretation, particularly for the other Website category. They were both included in the preferred model after a log-likelihood test confirmed that their inclusion made models 3 and 4 progressively better fits for the data. Model 4 provided the best fit and was used as the preferred model for this analysis.

Once these models have been estimated, they can be used to calculate expected probabilities for each category of the dependent variable. Three sets of calculations were made to determine expected probabilities of having arrived at the gateway page from each category of referrer for an ''average'' user at the mean values of each variable, for users from an on-campus IP address, and for users coming from an off-campus IP address. These probabilities are available in Table 2 and are depicted in Figure 1.

Results

Users with mean attributes seem, in general, to behave as expected. Regardless of resource type, they are more likely to access gateway pages through the AHSL Website. Although accesses through the UA Main Library Website are low overall, it is initially surprising to see such a difference in referrals to databases and referrals to other resource types. This difference can be explained by the inconsistency of catalog records for electronic journals and books in the main library catalog. Many electronic journals and books provided by the main library are not available in its catalog, let alone resources provided by another campus unit. Databases provided by AHSL, however, are frequently linked from the list of databases on the main library

Table 2
Expected probabilities of referrer by user type

	Database	Journal	Textbook
Average users			
P(AHSL Web)	0.8695	0.8631	0.9771
P(UA Main Library)	0.0521	0.0000	0.0080
P(Other UA Web)	0.0243	0.0095	0.0101
P(Search engine)	0.0517	0.1274	0.0048
P(Other Website)	0.0025	0.0000	0.0000
On-campus users			
P(AHSL Web)	0.9652	0.9894	0.9771
P(UA Main Library)	0.0021	0.0000	0.0080
P(Other UA Web)	0.0312	0.0075	0.0101
P(Search engine)	0.0014	0.0031	0.0048
P(Other Website)	0.0000	0.0000	0.0000
Off-campus users			
P(AHSL Web)	0.4288	0.3577	0.2569
P(UA Main Library)	0.0880	0.0002	0.0000
P(Other UA Web)	0.0119	0.0053	0.0020
P(Search engine)	0.4707	0.6365	0.7410
P(Other Website)	0.0006	0.0003	0.0001

Website, explaining the difference in referrals. The main library has, however, recently implemented an alphabetical list of all electronic journals available on campus, including those provided by AHSL, on its Website. This list may change these figures in the future.

The expected probabilities for users accessing the Website from on-campus are again much as expected. Surprisingly, the UA Main Library Website can be expected to be used less as a referrer to AHSL resources for on-campus users than for off-campus. This difference could be explained by on-campus users' greater familiarity with the libraries at UA, making them more likely to use the AHSL site for all health sciences–related research. The minimal use of external search engines is surprising given anecdotal evidence about user behavior, but it makes sense that on-campus users would take advantage of on-campus resources whenever possible.

Compared to on-campus users, fewer off-campus users arrived at AHSL gateway pages via on-campus tools, which was not surprising. It would be interesting, if the data were available, to isolate the usage of unaffiliated off-campus users (i.e., those who were not successfully authenticated to the AHSL proxy server) to see if this effect were strengthened. The probability of being referred from a search engine was higher for those seeking journals or textbooks, which confirmed the hypothesis that users seeking a known item would be more likely to use a search tool for navigation.

Overall, the AHSL Website appears to be the primary access tool for AHSL's primary users and an important access point for unaffiliated users. Although the data are crude and limited to Internet search engine use, users who might be looking for known items appear to use search engines more than those seeking another discovery tool. Because the expected probability of having been referred from within the AHSL Website is high, especially for on-campus but also for off-campus users, it would also be interesting to see what effect type of sought resource has on the mode of AHSL Website use. Mode of use is measured through the AHSLMODE variable, an unordered categorical variable with three categories: ALOE, Website browse, and Website search. ALOE is the library's catalog and contains records for all types of electronic resources, though the records are not comprehensive of what is available through the Website.

Figure 1
Probability of referrer by resource type

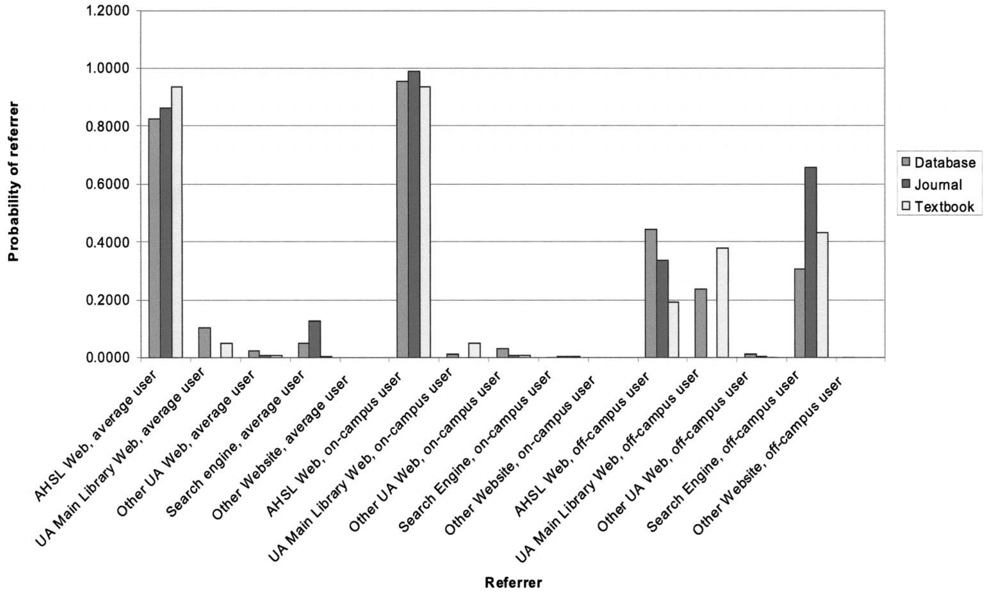


Table 3
MLR models of AHSL Website usage mode

Usage mode comparison	Model 1	Model 2	Model 3	Model 4
Browse Arizona Health Sciences Library (AHSL) Website/ALOE				
Constant	4.259563*	−9.163114*	−9.063193*	−10.13752*
UAIP	−1.998492*	2.542277*	2.396708*	3.561008*
JOURNAL		5.173898*	5.165612*	6.139324*
BOOK		2.068437*	2.061154*	2.136899*
ONCAMPUS		0.3610608†	0.2629466§	0.369915*
UAIPXONCAMPUS			0.1568758§	
UAIPXJOURNAL				−0.9922972§
UAIPXBOOK				
Search AHSL Website/ALOE				
Constant	2.495829*	−3.602031*	−3.562359*	−3.352156*
UAIP	−0.8487779*	1.568166*	1.496905*	1.206203*
JOURNAL		2.079569*	2.077467*	1.776099*
BOOK		−0.8723612*	−0.8754891*	−0.9360322*
ONCAMPUS		0.3185383*	0.2764696*	0.3435179*
UAIPXONCAMPUS			0.0809279§	
UAIPXJOURNAL				0.4389684*
UAIPXBOOK				
Pseudo R-Square	0.0560	0.1940	0.1940	0.1947
Log likelihood	−21137.622	−18047.979	−18047.493	−18032.322

* $P < 0.01$.
† $P < 0.05$.
‡ $P < 0.10$.
§ $P > 0.10$.

EFFECT OF RESOURCE TYPE ON MODE OF LIBRARY WEBSITE USAGE

Models

As with the previous set of models, it can be expected that users looking for a known item, such as a journal or a textbook, would want to access that item as quickly as possible and thus be more likely to be referred to a gateway page via a search tool than those seeking a tool for information discovery. In this case, two of the three usage modes are types of searches that should provide more detail for analysis. In each of these models (Table 3), the population is reduced to the segment that accessed the resource from a page in the AHSL Website (i.e., REFID = 1). Model 1 is a simple bivariate model in which mode of use (AHSLMODE) is regressed on user status (UAIP). In this model, a user's status (affiliated versus unaffiliated) has a significant effect on mode of Website usage for every category.

In a second model, the type of access restriction (UAONLY) and resource type (JOURNAL and BOOK) were added, all of which were significant across all categories of Website usage. An interaction between network location and access restriction (UAIPXONCAMPUS) was tested in model 3 but was non-significant. In model 4, the interaction between network location and journals (UAIPXJOURNAL) was tested and was significant for the search AHSL Website category. A log-likelihood test confirmed that this model was preferred to model 2.

The preferred model, model 4, was then transformed into expected probabilities of referral for users at means, on-campus users, and off-campus users. These probabilities are represented in Figure 2 and can be seen in Table 4.

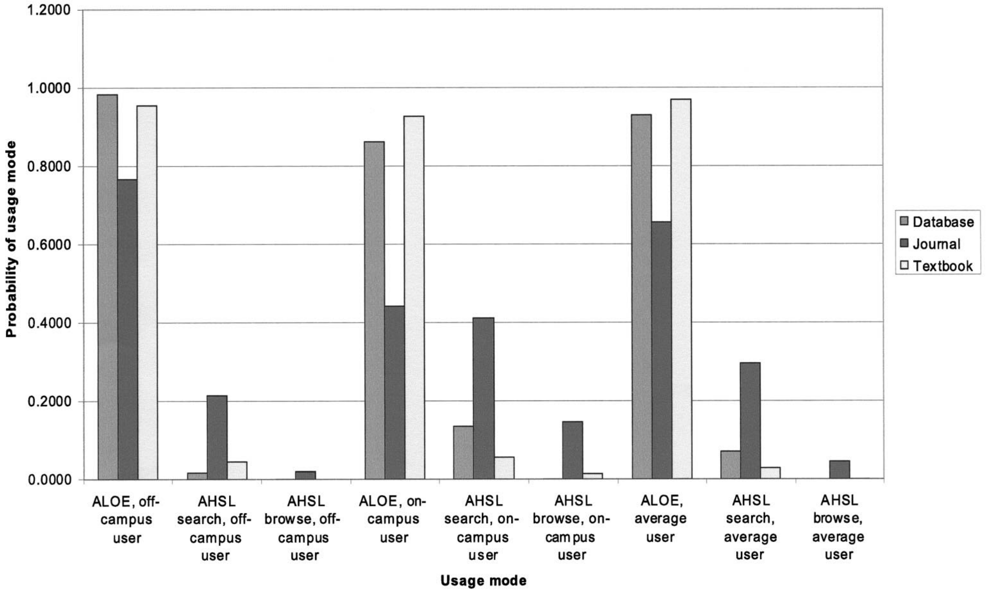
Results

Users with mean attributes seem to behave as expected when seeking databases and journals, which make up the largest portion of requests: Users are highly likely to browse to databases. The likelihood of searching for journals is higher, though users are still more likely to browse to journals than to search for them. Surprisingly, users are more likely to browse to textbooks than to databases.

Across resource type, on-campus users can be expected to search more than average users. This expectation is particularly true of users seeking journals, who are more likely to search the Website or ALOE than to browse to a journal. Interestingly, the probability of browsing to a textbook is still much greater than the probability of searching for one. This difference could indicate an age effect. Students might be more likely to use textbooks than other users and are likely to not need to use databases or journals due to the structured nature of the curricula in the health sciences. Students now have been exposed to the Web for much of their academic careers and may prefer browsing or searching the Web to the less flexible search of a library catalog.

The results for off-campus users are not surprising following the examination of probabilities for average and on-campus users. The most interesting point for these users is the extremely low expected probability of ALOE use across all resource types. While it is not surprising that these probabilities would be lower for

Figure 2
Probability of usage mode by resource type



off-campus users than for on-campus users, it could also be expected that, if the off-campus users were users of any library, some would be conditioned to use a catalog.

DISCUSSION

Understanding the navigation of electronic resources and the AHSL Website has implications for both system design and the information architecture of the Website. The understanding of user behavior gained through usage analysis can be used to improve the design of current systems, as Drott describes [21]. Information architecture has introduced libraries to structured approaches to Website development, a significant component of which involves understanding user behavior and designing sites that provide optimal navigation paths to content appropriate for a variety of information needs [22]. While designing such systems based on personal experience and anecdotal evidence may be possible in part, ideally design will also be based on empirical evidence.

Table 4
Expected probabilities of AHSL usage mode by user type

	Database	Journal	Textbook
Average users			
P(ALOE) at means	0.0004	0.0436	0.0030
P(AHSL search) at means	0.0631	0.3350	0.0273
P(AHSL browse) at means	0.9365	0.6215	0.9697
On-campus users			
P(ALOE)	0.0016	0.1192	0.0133
P(AHSL search)	0.1470	0.5109	0.0664
P(AHSL browse)	0.8515	0.3699	0.9204
Off-campus users			
P(ALOE)	0.0011	0.0193	0.0001
P(AHSL search)	0.0148	0.2192	0.0347
P(AHSL browse)	0.9841	0.7615	0.9651

Prior to this study, anecdotal evidence suggested many users navigate to electronic journals via Google. This study showed that while the probability that off-campus users access a journal page through any search engine is 0.66, the probability that an on-campus user does the same is only 0.003. So, if the site is really intended only for members of the library's primary user group, taking measures to enhance the discoverability of a gateway page in search engines will not be important. Alternatively, a library could consider this information evidence of an untapped market. For example, if the library provides a fee-based document-delivery service for external users, providing links to the service from gateway pages and then enhancing the visibility of the gateway pages in search engines could enhance the library's revenue stream.

Perhaps the simplest potential use of this sort of analysis would involve modifying the design of a Website, even a static one, to provide the most appropriate navigational tools for locating different information formats. Many library Websites are organized by information format, so better understanding of how users seek each format can point out problems in Website design and allow improvement of Website design. For example, from the findings in this study, search tools should be given as prominent a place as possible in the journals section of the site. Underused resources may need to be repositioned in a Website to make them easier to locate.

As described in the data mining literature, more sophisticated statistical analysis than that used in this analysis could be used to develop dynamic, predictive personalization systems that do not require specific user actions for customization. Several limitations in the library environment might make the development of such a system difficult, including the profession's concern with privacy and a lack of monetary, programming, and statistical resources. Nonetheless, developing library systems with such responsiveness as a goal could result in great improvements to their usability, even if the full vision is never realized.

The technique described in this paper does not provide a comprehensive methodology for analyzing Website usage, and the inferences that can be drawn from it are limited. Nonetheless, MLR and other regression techniques can provide a valuable supplement to log analysis. Even with a simple data set such as the one that was analyzed in this study, these techniques can provide some insight into usage not available through server logs. With additional preprocessing of Web transaction records and infrastructure that supports gathering data such as departmental affiliation, more valuable insights could be gained than were demonstrated in this paper. User and resource characteristics not available in logs may be analyzed, and these analyses may provide valuable insights into the behavior of users and the usage of the investments made by libraries on behalf of their constituents.

REFERENCES

1. Association of Research Libraries. Research library trends: an introduction. [Web document]. Washington, DC: The Association. [rev. 13 May 2003; cited 13 May 2003]. <http://www.arl.org/stats/arlstat/01pub/intro.html>.

2. Bertot JC, McClure CR, Moen WE, Rubin J. Web usage statistics: measurement issues and analytical techniques. Gov Inf Q 1997;14(4):373–95.

3. Hightower C, Sih J, Tilghman A. Recommendations for benchmarking Web site usage among academic libraries. Coll Res Libr 1998 Jan;59(1):61–79.

4. Haigh S, Megarity J. Measuring Web site usage: log file analysis. Netw Notes [serial online]. 1998;57 [cited 8 Apr 2003]. <http://www.nlc-bnc.ca/publications/1/pl-256-e.html>.

5. Stabin T, Owen I. Gathering usage statistics at an environmental health library Web site. Comput Libr 1997 Mar; 17(3):32–7.

6. Bauer K. Who goes there? measuring library Web site usage. Online 2000 Jan/Feb;24(1):30–1.

7. Breeding M. Monitoring the use of your Web site. Inf Today 2002 Dec;19(11):40.

8. Nicholas D, Huntington P, Lievesley N, Wasti A. Evaluating consumer Website logs: a case study of The Times/The Sunday Times Website. J Inf Sci 2000;26(6):399–411.

9. Li X. Library Web page usage: a statistical analysis. Bottom Line 1999;12(4):153–9.

10. Goldberg J. Why Web usage statistics are (worse than) meaningless. [Web document]. Self-pulished, 1995. [rev. 18 May 2001; cited 3 May 2003]. <http://www.goldmark.org/netrants/Webstats/>.

11. Kosala R, Blockeel H. Web mining research: a survey. ACM SIGKDD Explorations Newsl 2000 Jun;2(1):1–15.

12. Dunham MH. Data mining: introductory and advanced topics. Upper Saddle River, NJ: Pearson Education, 2003.

13. Han J, Kamber M. Data mining: concepts and techniques. San Diego, CA: Academic Press, 2001.

14. Feng T, Murtagh F. Towards knowledge discovery from WWW log data. In: Proceedings of the IEEE International Performance, Computing, and Communications Conference. Phoenix, AZ: February 2000:302–7.

15. Davis P. Information-seeking behavior of chemists: a transaction log analysis of referral URLs. J Am Soc Inf Sci Tech 2004;55(4):326–32.

16. Papatheodorou C, Kapidakis S, Sfakakis M, Vassiliou A. Mining user communities in digital libraries. Inf Tech Libr 2003 Dec;22(4):152–7.

17. Wormell I. Matching subject portals with the research environment. Inf Tech Libr 2003 Dec;22(4):158–64.

18. Geyer-Schulz A, Neumann A, Thede A. An architecture for behavior-based library recommender systems. Inf Tech Libr 2003 Dec;22(4):165–74.

19. Zucca J. Traces in the clickstream: early work on a management information repository at the University of Pennsylvania. Inf Tech Libr 2003 Dec;22(4):175–8.

20. Long JS. Regression models for categorical and limited dependent variables. Thousand Oaks, CA: Sage, 1997.

21. Drott MC. Using Web server logs to improve site design. In: Proceedings of the 1998 ACM International Conference on Systems Documentation, Association for Computing Machines, 1998:43–50.

22. Rosenfeld L, Morville P. Information architecture for the World Wide Web. 2nd ed. Sebastapol, CA: O'Reilly and Associates, 2002.

Received June 2003; accepted April 2004